CS 6322: Information Retrieval

**Homework 1**

**Problem 1 - Tokenization**

**Ans 1 –** The output for problem 1 (Tokenization) is given below:

1. There are 201520 number of tokens in the Cranfield text collection.
2. There are 8514 number of unique words in the Cranfield text collection.
3. Total of 3551 words occur only once in the Cranfield text collection.
4. Top 30 most frequent words in the Cranfield text collection -

[('the', 18682), ('of', 11312), ('and', 5693), ('a', 5283), ('to', 4372), ('in', 4235), ('is', 4107), ('for', 3265), ('are', 2425), ('with', 2082), ('by', 1695), ('at', 1591), ('that', 1565), ('on', 1552), ('flow', 1412), ('be', 1269), ('an', 1258), ('as', 1101), ('this', 1079), ('from', 1067), ('pressure', 1021), ('which', 968), ('number', 905), ('results', 873), ('it', 851), ('mach', 725), ('boundary', 722), ('was', 698), ('theory', 693), ('method', 634)].

1. There are 143.94 average number of word-tokens per document.

**Program Description -**

    Initially entire Cranfield text collection is loaded in the memory and iterated over each document in the collection. Xml-DOM(Document Object Model) API is used to parse the SGML document in the collection as the format of each document is in the SGML. It will fetch the text part of the document and will ignore the tags. Any punctuation and numeric in the document are removed by the regex (regular expression). After that the text is split by space and a tokens list is used for saving it.

1. The time taken by program is around 2.09 seconds for execution and the output for tokenization and stemming is produced by the program.
2. Tasks performed by the program:
3. Conversion of all the texts into lower case.
4. Deleting the dashes from the words and merging them.
5. All the possessives from the words are deleted through regex.
6. It also removes the acronyms.
7. Data structures used:

a) Produced tokens are stored in list.

b) Distinct tokens are stored in set.

**Problem 2 - Stemming**

**Ans 2 -**   The output for problem 2 (Stemming) is given below -

1. There are 5839 number of distinct stems in the Cranfield text collection.
2. There are 2428 number of stems with single occurrence in the Cranfield text collection.
3. Top 30 most frequent stems in the Cranfield text collection -

[('the', 18682), ('of', 11312), ('and', 5693), ('a', 5283), ('to', 4372), ('in', 4235), ('is', 4107), ('for', 3265), ('ar', 2426), ('with', 2082), ('on', 1818), ('by', 1695), ('flow', 1598), ('at', 1591), ('that', 1565), ('be', 1366), ('an', 1258), ('number', 1236), ('pressur', 1180), ('as', 1101), ('thi', 1079), ('result', 1072), ('from', 1067), ('it', 1029), ('which', 968), ('effect', 839), ('method', 820), ('solut', 785), ('theori', 780), ('boundari', 750)]

1. There are 143.94 average number of word-stems per document.

**Description** -

* For stemming the tokens produced by the program, an open-source porter stemmer <https://tartarus.org/martin/PorterStemmer/> is used.
* For producing the stems from tokens, PorterStemmer class is used as the utility in the main class.
* In the folder, one README file is attached that describes steps to execute the program.